Reviewer’s report

Title: ‘Caveat emptor’: the cautionary tale of endocarditis and the potential pitfalls of clinical coding data: an electronic health records study

Version: 0 Date: 11 Mar 2019

Reviewer: Joan Casey

Reviewer's report:

Summary
The authors review data from Leeds Teaching Hospital and Oxford University Hospital Trust to compare the efficacy of diagnostic codes and objective criteria used to define gold standard infective clinical endocarditis cases. They find the use of any endocarditis code overestimates the true incidence of endocarditis and suggest several ways to improve identification of true endocarditis cases (i.e., removing specific codes, short stays, readmissions). They focus on the use of ICD codes. I found several aspects of the manuscript difficult to follow, likely because so much material is relegated to the appendix. I list several questions and suggestions below.

Major
1. It is unclear from lines 112-117 which codes the authors selected how Table S3-S4 work together to make up the definition. As this is a critical piece of the manuscript, I suggest being more explicit and potentially moving Table S3 to the main manuscript.
2. Lines 178-183: why were readmissions classified as occurring within 30 days of a prior endocarditis admission? I see this information in the appendix, but a bit more is needed in the methods, particularly "we excluded coded admissions for <3 days where the patient was discharged alive." What was the average length of stay for an endocarditis patient? Wouldn't this definition skew towards patients with less severe or less-well diagnosed endocarditis? Would it make more sense to classify readmission based on a timeframe after the discharge date?
3. The methods state that incidence trends were estimated with annual counts and population data for 2001-2016. Did the denominator (i.e., the population count offset) change every year or was this stable?
4. I was confused by the term "admission." After reading the NHS website and definitions, I see a finished admission episode is "the first period of inpatient care under one consultant within one healthcare provider." I would include this or however "admission" is defined in this manuscript in the methods section. Also, some explanation for the strict focus on admission rather than hospital stay is of interest (unless this is untrue). Along these lines, the section in lines 240-247 is difficult to follow. If you are only looking at codes during the admission and someone developed infective endocarditis during the admission or during chemo, how would this code appear at the time of admission? Then in section 259-265 it sounds like codes/diagnostic criteria from throughout the hospital stay are used: "developing infective endocarditis as inpatients..." This may be a lack of my understanding about how these admission codes are assigned and the timeframe that encompasses the admission. What is the
timeframe during which a code would be considered an admission code? Are they retrospectively assigned? What was the average length of admissions in the study? The timing of codes and other information used to identify endocarditis in the analysis needs to be clarified in the methods/results.

5. Lines 272-276: were any other methods tried to improve PPV other than those listed? Did some attempted methods not work? How were these particular methods selected?

6. Line 303: were streptococcal cases differentially newly reported in the secondary position relative to other organisms? Or did all organisms start getting reported in this position over time? It's unclear to me how this would translate into overestimation of streptococcal cases if researchers appropriately deal with "no organism reported" vs. a specific organism reported. If endocarditis cases were not tagged with an organism earlier in the study period, wasn't that under-reporting? Do you think that the actual causative organisms have changed over time? The section where you report on causative organisms could be much clearer. (1) What were the causative organisms? (2) How were these tested for/reported/with what frequency? (3) Changes in testing/reporting over time. How were organism-specific trends estimated? If fewer cases had an accompanying organism code in the past, would it not make sense to include an offset for "any identified organism" or to otherwise control for non-reporting?

7. It might be useful to discuss the concept of electronic phenotyping and rule-based vs. probabilistic identification of disease. Some conditions have consensus definitions from the EHR, see, for example, https://phekb.org/phenotypes. Why limit yourself to diagnostic codes since their use likely varies across healthcare systems. A major strength of EHR data is the huge amount of information, beyond codes, that is present. The discussion could at least mention these possibilities. An example of this process is described here: Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, Sweeney TE, Gyang E, Shah NH. Learning statistical models of phenotypes using noisy labeled training data. Journal of the American Medical Informatics Association. 2016 May 12;23(6):1166-73.

Minor

1. It would be clearer to always refer to "clinical cases" as "confirmed clinical cases."
2. Line 117: Please explicitly state why codes related to viral endocarditis were removed.
3. Lines 131-133: please rephrase there is either a typo or this sentence is simply hard to follow.
4. Line 146: "all diagnostic codes from all consultant episodes." Over what time period? The entire hospital stay?
5. It would be helpful to break out the information about "causative organism" into a separate paragraph in the methods. Please list the potential organisms and describe a bit about the standard procedure for testing, including AMR. Then, in the results, line 294 please add a broader statement about the common causative organisms, i.e., majority Streptococcus and Staph. In addition, how accurate do you think the secondary Streptococcus codes were, did you have access to laboratory testing data at Leeds? Leeds had clinician-reported organism?
6. Line 154: please provide actual Duke criteria in the text
7. Line 161: how was this order selected? Is this standard?
8. Line 158-159: "...code was present in any consultant episode, in any position..." Figure 1 shows primary and secondary codes. How many codes were available per encounter? Up to 5? 10? Please specify in text.
9. Figure 2: Please use an alternative color scheme, many people are red-green colorblind.
10. Line 202: which Figure?
11. Line 259: this information comes as a surprise. Please add the fuzzy text search methods to the methods section.
12. Line 286 and other places, recommend removing "did not reach statistical significance." Readers can look at the confidence intervals and draw conclusions. If the authors choose to refer to "statistical significance" throughout, please add a statement about which alpha level they refer to in the
methods section.
13. Figure 3: resolution is so poor that I cannot read the legend, which might provide this information, but the caption does not. How were these lines fit? What is the grey? The 95% CI? Please add model details to the figure caption.
14. The use of lines in Figure S4 is confusing as these categories are not linked over time.

**Are the methods appropriate and well described?**
If not, please specify what is required in your comments to the authors.

No

**Does the work include the necessary controls?**
If not, please specify which controls are required in your comments to the authors.

Not applicable

**Are the conclusions drawn adequately supported by the data shown?**
If not, please explain in your comments to the authors.

No

**Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?**
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

**Quality of written English**
Please indicate the quality of language in the manuscript:

Acceptable

**Declaration of competing interests**
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests.

**Statement on potential review bias**
Please complete a statement on potential review bias, considering the following questions:

1. Did you co-author any publication with an author of this manuscript in the last 5 years?

2. Are you currently or recently affiliated at the same institution as an author of this manuscript?

If you can answer no to all of the above, write 'I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria'. If your reply is yes to any, please give details below.

I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license ([http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/)). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal